Identifikacija spletno specifičnih kolokacij pogostega besedišča

Senja Pollak

Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
senja.pollak@ijs.si

Povzetek

V prispevku predstavimo pristop k luščenju spletno specifičnih kolokacij pogostega slovenskega besedišča. Izbrali smo več kot 150 najpogostejših besed, ki se pojavljajo tako v korpusu uporabniških vsebin Janes kot v vzorčenem referenčnem korpusu slovenščine Kres. Za te leme smo izluščili kolokatorje, ki se pojavljajo tik pred izbrano besedo in izluščili sezname za oba korpusa. Nato smo identificirali tiste kolokacije, ki se pojavljajo izključno v korpusu uporabniških vsebin. Predstavljena metodologija omogoča hitro luščenje in modularne nastavitve parametrov, primernih za luščenje podobnih specifičnih seznamov kolokacijskih kandidatov.

Identification of Web-specific collocations of common lexis

In this article we present an approach to extracting Web-specific collocations for standard Slovene lexis. We selected more than 150 most frequent Slovene words in Janes, the corpus of user-generated content, and Kres, the reference balanced corpus of Slovene. In each of the two corpora we extracted collocators which immediately preceded the selected words. We then identified the collocations that were specific to user-generated content, i.e. those appearing only in the Janes corpus. The presented methodology enables a rapid extraction of collocation candidates and modular parameter settings, suitable for extracting similar lists of specific collocation candidates.

1 Uvod

Kolokacije so tipične sopojavitve besed. Pristopi k preučevanju kolokaciji se v grobem delijo na frekvenčne in frazeološke (Nesselhauf, 2005). *Frekvenčni pristopi* razumejo kolokacije kot statistične sopojavitve besed znotraj določenega okna[1] (prim. Firth, 1957; Halliday, 1961; Sinclair, 1966), tipičnost oz. statistična signifikantnost sopojavitev pa je definirana s pogostostjo in različnimi merami statistične povezanosti. Iz te tradicije, ki razume kolokacije v širšem pomenu, izhajajo tudi korpusni pristopi in računalniški pristopi k luščenju kolokacij, sama dolžina okna (definicija sopojavitve), frekvence in statistične mere povezanosti pa se pojavljajo v več različicah. *Frazeološki pristopi* pa obravnavajo kolokacije v ožjem pomenu. Poudarjajo sintaktične (Sinclair, 1991; Kjellmer, 1987), predvsem pa semantične aspekte kolokacij, ki jih definirajo kot besedne zveze, ki se nahajajo med »prostimi« besednimi kombinacijami in idiomi (npr. Cowie, 1994; Benson, 1989). Razlikovalni kriteriji pa se opirajo predvsem na pomensko razstavljivost in transparentnost zveze. Avtorji se tudi razlikujejo po tem, ali se osredotočajo le na leksikalne besede ali tudi na slovnične besede (npr. Bartsch, 2004; Siepmann, 2005).

V prispevku se pri postopku luščenja naslanjamo na strogo statistične kriterije. Z začetnim izborom lem, kjer upoštevamo izključno samostalnike, ter pri podajanju primerov, kjer diskutiramo tako o slovničnih kot o pomenskih strukturah, pa se odmikamo od strogo frekvenčnega pristopa.

Kolokacije lahko razumemo kot pare besed (predvsem pri frekvenčnih pristopih) ali pa izbrano besedo (lemo) interpretiramo kot bazo, ki ga kolokator pobliže določa (Hausmann,1989 po Gorjanc in Jurko, 2004). V tem prispevku luščimo kolokacije kot pare samostalniške leme *l* in kolokatorja *k*.

Kolokacije so pomembno področje preučevanja jezika in tako tudi za analizo spletne slovenščine. Termin *spletni jezik* oz. *jezik uporabniških vsebin* uporabljamo za jezik komunikacije na forumih, blogih in družbenih omrežjih, kot je Twitter. Omejujemo se na preučevanje točno določenega korpusa, korpusa Janes (Fišer et al., 2014), tako da ne pokrivamo celotnega spektra spletnega jezika (npr. spletnega časopisja, komunikacije preko elektronske pošte). Prav tako v naši obravnavi izpuščamo besedila nekaterih pomembnih družbenih omrežij (npr. Facebook) in marsikatere uporabniške vsebine (npr. Wikipedija). Zato mora bralec izraz *spletni jezik*, ki ga uporabljamo v nadaljevanju, primerno interpretirati.

Spletni jezik ima svoje značilnosti. Pisno spletno komunikacijo določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki (Noblia, 1998; Fišer et al., 2014). Bolj kot je izbrana oblika komuniciranja interaktivna, poteka v realnem času in ima na drugi strani prisotnega sogovornika, več prvin spontanega govorjenega jezika vsebuje, vključno s prozodičnimi elementi kot tudi s paralingvističnimi elementi, prilagojenimi za računalniško komunikacijo (Crystal, 2001). Specifike spletnega jezika lahko preučujemo z različnih vidikov, kot je ortografija (npr. bolj fonetičen zapis besed, opuščanje ločil in ponavljanje črk za čustveno poudarjanje zapisane izjave), skladnja (npr. nestandardni vrstni red, stavčne strukture), diskurzivna raven (interaktivnost) in leksikalna raven, kjer se najhitreje in najpogosteje kaže inovativna raba jezika, značilna za uporabniške vsebine (neologizmi, aktualne tematike). V tem prispevku nas zanima predvsem leksikalna specifika, ki jo preučujemo na nivoju večbesednih zvez oz. kolokacij.

V članku predstavimo metodo identifikacije kolokacij, ki so specifične za spletne vsebine, ali natančneje, za identifikacijo tistih kolokacij, ki se pojavljajo le v jeziku

[1] Za razliko od besednih nizov oz. n-gramov, ki so zaporedja besed, okno upošteva sopojavitev besed, četudi ne gre za pravo zaporedje. Npr. v nizu besed b_1 b_2 b_3 je pri pri preučevanju kolokacije besede b_1 in oknu 2 upoštevana tudi beseda b_3. Pri oknu 1 oz. bigramih $b_1 b_2$ pa ni razlike.

(oz. korpusu) spletnih uporabniških vsebin in ne v bolj splošni, »standardni« rabi jezika, ki jo predstavlja referenčni korpus. Za izhodiščne leme (tj. leme, za katere nas zanimajo njihova kolokacijska okolja) izberemo najpogostejše splošno (torej nespecifično) slovensko besedišče.

S področja kolokacij je na voljo vrsta tujih študij. Poleg osrednjih leksikografskih vidikov (Castro in Faber, 2014) so kolokacije pomembne z vidika uporabe v številnih aplikacijah, kot so poučevanje tujega jezika (Orenha-Ottaiano, 2012), luščenje informacij (Lin, 1998), strojno prevajanje (Gerber in Yang, 1997) itn. Na uporabniške vsebine se osredotočajo npr. Rösiger et al. (2015), ki luščijo terminologijo iz korpusa uporabniških vsebin strani tipa »naredi si sam«), Seretan (2015) se posveča prevodom večbesednih izrazov iz uporabniških spletnih vsebinah z vidika strojnega prevajanja, analizo sentimenta v uporabniških spletnih vsebinah obravnava npr. Yu (2014).

V slovenščini se je z luščenjem kolokacij ukvarjalo že več avtorjev. Za leksikografske namene so podatke iz referenčnega korpusa Gigafida luščili Gantar in Krek (2011), Kosem et al. (2013), s terminološkega vidika so bile kolokacije obravnavane v Vintar (2010) in Logar Berginc et al. (2014).

V dosedanjem delu smo že obravnavali tematiko iskanja kolokacij, specifičnih za korpus spletnih vsebin (Pollak, 2015), vendar se od preteklih raziskav pričujoči prispevek razlikuje v več pogledih:

- luščimo spletno specifične kolokacije lem najpogostejšega splošnega besedišča,
- nova metodologija vključuje orodje za hitri izvoz kolokacijskih seznamov (API),
- v prejšnjem eksperimentu smo se osredotočili na analizo okoli 30 lem, v tem eksperimentu pa gre za mnogo večji izbor, ki obsega 150 lem,
- modularna nastavitev parametrov (frekvenca, mera kolokabilnosti, mera razpršenosti kolokatorja) omogočajo enostavno uporabo orodja glede na specifične cilje.

2 Metodologija luščenja

Za poljubno lemo *l,* ki se pojavlja v seznamu *n* najpogostejših besed slovenskega besedišča (tako spletnega kot splošnega) nas zanimajo tiste kolokacije oz. pari kolokatorjev in lem <*k*, *l*>, ki se pojavljajo izključno v uporabniških spletnih vsebinah, ne pa v referenčnem korpusu slovenščine. Z dodatnimi parametri definiramo ustrezno mero kolokabilnosti, frekvence v korpusu, okno iskanja ter razpršenost kolokatorja.

2.1 Korpusa

Za iskanje razlik med kolokacijami slovenščine uporabniških spletnih vsebin in »splošno« slovenščino uporabimo dva korpusa. Spletno slovenščino predstavlja korpus Janes v0.3 (Fišer et al., 2014), ki zajema besedila forumov, tvitov in blogov (okoli 161 milijonov pojavnic).

Kot referenčni korpus »splošne« slovenščine izbremo korpus Kres (Logar et al., 2012). Kres je iz korpusa Gigafida (ibid.) vzorčeni uravnoteženi podkorpus in ga uporabljamo kot referenčni korpus. Ima 121 milijonov pojavnic in vsebuje stvarna besedila, leposlovje, časopisje, revije in internetne vsebine.

Metodologija je neodvisna od izbora korpusov, vendar korpus definira same rezultate. Če bi vzeli drug referenčni korpus (npr. Gigafida), bi bile tudi izluščene spletno specifične kolokacije druge. V primeru križanja seznamov z govornim korpusom (Verdonik et al., 2014) pa bi dobili le tiste kandidate, ki so specifični za uporabniške vsebine, ne pa za govor. V primeru sproti posodabljanega korpusa »splošne« oz. standardne slovenščine pa bi z metodo lahko natančneje luščili spletno oz. nestandardno besedišče.

2.2 Izbor lem

V predstavljeni raziskavi nas zanimajo spletno specifične kolokacije za leme splošnega slovenskega besedišča. Izdelali smo frekvenčne sezname lem po posameznih besednih vrstah. Za opisano raziskavo smo se osredotočili na 250 najpogostejših samostalnikov vsakega korpusa. Seznama iz obeh korpusov smo med seboj križali in ohranili 150 občnih samostalnikov, ki se pojavljajo na obeh seznamih.

Nekaj samostalnikov iz seznama (izbor je naključen in zajema primer vsake črke abecede): *avto*, *beseda*, *cerkev*, *človek*, *dogodek*, *energija*, *fant*, *glava*, *hiša*, *igralec*, *jezik*, *knjiga*, *ljubezen*, *mati*, *namen*, *oče*, *pesem*, *roka*, *sistem*, *šola*, *telo*, *ura*, *vas*, *zgodba*, *želja.* Primeri tistih lem, ki so bili na seznamu le enega od obeh korpusov in jih zato nismo vključili, so npr. leme iz pravnih besedilih korpusa Kres (*ministrstvo, komisija, člen, besedilo*) ter npr. leme korpusa Janes, vezane na spletni medij (*novica, video, forum, komentar*).

Luščenje spletno specifičnih kolokacij lem pogostega splošnega slovenskega besedišča, ki je opisano v tem prispevku, je predvsem koristno za preučevanje pomenskih premikov, iskanje aktualnih kolokatorjev splošnega slovenskega besedišča ipd. Za razliko od te naloge bi lahko preučevali tudi leme, ki se pojavljajo izključno v korpusu Janes in ne v referenčnem korpusu, vendar v tem primeru ni treba primerjati kolokatorjev z referenčnim korpusom. Prav tako zanimiva je naloga preučevanja zelo pogostih besed v korpusu Janes, ki so v referenčnem korpusu manj pogoste, vendar tu najverjetneje ne bi želeli izključiti kolokacij, ki se pojavljajo v referenčnem korpusu (npr. *pisanje bloga* se kot kolokacija leme *blog* pojavlja v obeh korpusih), temveč bi rajši uporabili mero, ki primerja moč kolokacij (prim. Pollak in Arhar Holdt, 2015; Pollak, 2015).

Izbor lem je torej usklajen z motivacijo članka (iskanje nestandardnih in novih kolokacij zelo pogostega splošnega besedišča) in z metodologijo iskanja kolokacij (izključnost pojavljanja v korpusu uporabniških spletnih vsebin). Druge vidike bomo podrobneje preučili v nadaljnjem delu, možno metodologijo za njihovo obravnavo pa smo že predstavili v Pollak in Arhar Holdt (2015) in Pollak (2015).

2.3 Luščenje kolokacij iz posameznih korpusov

Na ravni izbora lem smo upoštevali leme, ki se pojavljajo v obeh korpusih, pri luščenju kolokacij pa je cilj izluščiti samo tiste kolokacije (kombinacije lem in kolokatorjev), ki so izključno v spletnih uporabniških vsebinah (glej 2.4), vendar moramo zato najprej izvoziti kolokacijske sezname iz obeh korpusov.

Za luščenje kolokacij uporabljamo funkcijo *kolokacije* orodja SketchEngine (Kilgariff et al., 2004). Hitro

luščenje za veliko število izbranih lem ter poljubno nastavljanje paramerov smo omogočili z API-ja, ki kliče lokalno instalacijo korpusov.

Nastavitve parametrov luščenja kolokacij v orodju Sketch Engine zajemajo razpon okna, tj. okolico besed izhodiščne leme, mero za izračun kolokacijske vrednosti ter minimalni frekvenci leme in kolokacijskega niza v izbranem korpusu.

V predstavljenih eksperimentih smo se omejili le na besede tik pred izbrano lemo, torej na okno -1, 0, vendar metoda sama ni odvisna od nastavitve okna.[2]

Za izračun kolokacijske vrednosti kolokacij za posamezni korpus uporabimo mero logDice (Rychlý, 2008), ki je priporočena mera za izračun kolokacij v orodju SketchEngine:

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

D se v formuli nanaša na originalni Diceov koeficient (Dice, 1945), medtem ko se f_x, f_y in f_{xy} nanašajo na relativne frekvence besed *x* in *y* ter njunih skupnih pojavitev. Ta mera ima več prednosti (Rychlý, 2008), kot so lahka interpretacija vrednosti (med 0 in 14) ter neobčutljivost na velikosti korpusov, saj temelji na relativnih frekvencah in omogoča primerljivost med različnimi korpusi.

Za minimalne frekvence luščenja smo pri luščenju kolokacij posameznega korpusa določili nastavitev 5 pojavitev za kolokator in 3 za kolokacijski niz v izbranem oknu, vendar se v koraku izbora spletno specifičnih kolokacij (glej 2.4) omejimo na kandidate z bolj pogostimi pojavitvami (min. frekvenca 10).

Za vsako lemo izvozimo kolokatorje in pripadajoče logDice vrednosti ter relativne frekvence.

V Tabeli 1 prikažemo najmočnejše kolokatorje leme *družina*, ki jo izberemo za ponazoritev postopka luščenja tudi v nadaljevanju prispevka.

Kolokacije (lema: *družina*)	
Kres	**Janes**
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(*n*) članska družina
svoja družina	romska družina

Tabela 1: Najpogostejših deset kolokacij z besedo družina v korpusih Kres in Janes (nastavitve: -1, 0, mera logDice).

Vidimo, da je več kolokacij presečnih, vendar se v kolokacijah korpusa Janes bolj odražajo aktualne tematike (npr. kolokacija *enostarševska družina* se pojavlja tudi v korpusu Kres, vendar je logDice vrednost bistveno nižja (5,7 za Kres; 8 za Janes), kakor tudi število pojavnic (63 v korpusu Kres oz. 0,52/1M, 199 oz. 1,23/M v korpusu Janes). Podobno velja za nekatere druge tematsko aktualne primere, npr. *primarna družina* ima logDice vrednosti 7,6 v korpusu Janes (156 pojavnic) in 4,8 v korpusu Kres (34 pojavnic), kar pomeni, da je tematika veliko bolj zastopana v spletnih vsebinah, kljub temu da kolokacija obstaja v obeh korpusih. V našem prispevku se za razliko od Pollak (2015) osredotočimo na kolokacije, ki se pojavljajo izključno v enem od dveh korpusov, in tovrstnih primerov, kjer gre le za razliko med vrednostma kolokatorjev, podrobneje ne obravnavamo, kljub temu da predstavljajo zanimiv material za analizo.

2.4 Identifikacija spletno specifičnih kolokacij

Spletno specifične kolokacije iščemo s pomočjo programa, ki ga izdelamo v ta namen. Ta na podlagi izvoženih kolokacijskih seznamov dveh korpusov izlušči seznam tistih kolokacij oz. kolokatorjev dane leme, ki se pojavljajo izključno[3] v izbranem korpusu (v našem primeru v korpusu Janes).[4] Poleg ključnega pogoja iskanja kolokacij, ki se pojavljajo izključno v spletnem korpusu, omogočamo tudi nastavitev najnižje dopuščene frekvence kolokacije v korpusu, najnižje vrednosti logDice ter razpršenost kolokatorja po lemah (v nadaljevanju *razp*).

Razp je odstotek lem, ki vsebujejo določeni kolokator.[5] *Razp* izračunamo tako, da število izdelanih kolokacijskih seznamov specifičnega korpusa, ki vsebujejo določeni kolokator, delimo s številom vseh preučevanih seznamov kolokacij tega korpusa (tj. s številom obravnavanih lem). Torej, če se nek kolokator pojavlja na kolokacijskih seznamih vseh preučevanih lem, je njegova vrednost 1 in čim redkeje se pojavlja, tem bolj je specifičen za določeno lemo in vrednost *razp* je bližje 0. S to mero lahko določimo zgornjo mejno vrednost in izločimo funkcijske besede z visoko *razp* vrednostjo, ki se pojavljajo kot kolokatorji velikega števila lem (npr. *vsak, kakšen, moj*)[6]. Tem je sicer pripisana praviloma nizka vrednost logDice. Prav tako imajo relativno visoko *razp* vrednost pogoste besede, ki so specifika korpusa, npr. besede z opuščenimi strešicami (*vec*), pogoste okrajšave v nestandardnem zapisu (*slo*), pogosti tematski izrazi (*političen*), razlike med lematizaroji dveh korpusov (edini vs. edin). itd.

[2] V Pollak (2015) smo npr. vzeli okno -3 +3.

[3] To pomeni, da se v referenčnem korpusu pojavljajo manj kot trikrat, saj je bil ta pogoj upoštevan pri luščenju iz posamičnih korpusov (glej sekcijo 2.3).

[4] Alternativni pristop k iskanju korpusnospecifičnih kolokacij je, da se ne omejimo na kolokatorje, ki se pojavljajo izključno v izbranem korpusu, temveč dopuščamo njihove pojavitve v obeh korpusih, vendar določimo razliko v vrednosti kolokabilnosti (mero CorpDiff smo vpeljali v Pollak in Arhar Holdt (2015) in jo uporabili v Pollak (2015)). V pričujoči raziskavi specifičnost definiramo kot izključnost.

[5] *Razp* ima enako motivacijo kot mera *inverse document frequency* oz. *idf*, ki jo je uvedla K. Spärck Jones (1972) v kontekstu iskanja dokumentov (angl. *document retrieval*). Mera *idf* uteži termine tako, da zmanjša vrednost terminov, ki se pojavljajo v več dokumentih neke zbirke, in poveča vrednost tistim, ki so prisotni v manj dokumentih. V našem primeru uporabimo različico *frekvence po dokumentih* (angl. *document frequency*) in ne *idf,* zato je manjša vrednost bolj informativna.

[6] Kolokatorji z vrednostjo *razp*, ki presega 0,9, so: *a, ampak, brez, dober, en, glede, isti, kak, kako, kakšen, moj, morati*, *nek*, *nekaj*, *niti*, *njihov*, *noben*, *oz.*, *reči*, *svoj*, *tak*, *tisti*, *torej*, *tvoj*, *vaš*, *veliko*, *vsak*, *zaradi.*

Kolokator	Lema	Frekvenca	Rel. frek.	LogDice	Razp.
festival	družina	81	0,632	6,313	0,093
celje	družina	44	0,344	5,073	0,258
homoseksualen	družina	21	0,164	4,783	0,066
ožji	družina	19	0,148	4,638	0,040
narcisističen	družina	16	0,125	4,431	0,007
janšev	družina	18	0,141	4,314	0,205
prazničen	družina	14	0,109	4,088	0,152
razdreti	družina	10	0,078	3,715	0,033
razpasti	družina	10	0,078	3,611	0,046
priloga	družina	8	0,062	3,313	0,033

Tabela 2: Izpis 10 najvišje rangiranih specifičnih kolokacijskih kandidatov za lemo *družina* (mera logDice).

Čeprav je preučevanje kolokatorjev z relativno visoko vrednostjo *razp* zanimivo z vidika razumevanja spletnega diskurza (npr. kolokatorji *sploh*, *pač, tale*) in nestandardne uporabe leksike (*rabiti* v pomenu *potrebovati*), pomenskih premikov na ravni besed (*hud* v pomenu *dober*) ali zapisa (npr. kratica *slo*), so bolj zanimivi za samostojno preučevanje ali v kombinaciji z njihovi kolokatorji kot z vidika preučevanja kolokacij drugih lem.

Izdelani končni seznami kolokacij uporabniških vsebin vsebujejo par <kolokator, lema>, ki mu pripišemo še dodatne informacije, kot so frekvenca kolokacijskega niza, relativna frekvenca (glede na milijon besed v korpusu Janes), logDice vrednost, kolokatorja z vrednostjo *razp* ter povezavo do konkordanc v korpusu. Končna oblika izpisa je prikazana v Tabeli 2, dodane pa so ji še povezave na konkordance iz korpusa, kar pa zaradi predolgih povezav v prispevku izpuščamo.

3 Rezultati raziskave

Predstavljena metodologija omogoča izdelavo različnih seznamov s poljubnimi nastavitvami (okno, mera kolokabilnosti, minimalna frekvenca, minimalna vrednost mere kolokabilnosti, uporaba statistike *razp*, izključnost[7]). Kot je podrobneje opisano v prejšnjem poglavju v naši raziskavi izberemo parameter za dolžino okna -1, 0, mero logDice in parameter izključnosti. Od nastavitev je odvisna tudi količina izluščenih kandidatov. V Tabeli 3 prikažemo razliko v številu izluščenih kandidatov glede na uporabo različnih nastavitev mejne vrednosti logDice, minimalne frekvence ter uporabe filtriranja z mero *razp.*

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Tabela 3: Število izluščenih kolokacijskih kandidatov, ki se pojavljajo v korpusu Janes in ne v korpusu Kres, za 151 izbranih lem, z različnimi nastavitvami parametrov.

V sorodni študiji (Pollak, 2015) smo glavne kategorije izluščenih kolokacij ločili na aktualno tematiko, spletno tematiko, frazeologijo, lastna imena in specifične kolokatorje. V tem prispevku se osredotočimo predvsem na metodologijo luščenja, za razliko od Pollak (2015) pa iščemo izključno spletne kolokacije splošnega besedišča. Za ponazoritev različnih nastavitev pa se osredotočimo na lemo *družina.*

Strožji kot so kriteriji (glej Tabelo 3), manj kandidatov dobimo v pregled. Na primeru *družina* nam z najstrožjimi nastavitvami iz Tabele 3 (zadnja vrstica) ostanejo naslednje kolokacije:

- *homoseksualna*, *narcisistična*, *ožja družina*
- *razdreti družino*
- *Festival družin*

Še en kandidat, in sicer par <razpasti, družina>, je le delno relevanten, saj polovica konkordanc izhaja iz samostalniške besedne zveze *razpadla družina,* ki pa se dejansko pojavlja v obeh korpusih. Ob iskanju zaporedne sopojavitve lem *razpadel* (pridevnik) in *družina*, najdemo kolokacijo *razpadla družina* v korpusu Kres štirikrat, v korpusu Janes pa trikrat. Kolokacijski par, ki je izluščen kot specifičen za korpus Janes, pa izhaja iz glagola *razpasti* in besede *družina,* vendar je od desetih primerov le v štirih dejansko uporabljen glagol *razpasti,* v šestih pa

[7] Izključnost pomeni, da se kolokacija pojavlja le v enem od dveh korpusov, kar je tudi izbrani pristop v tem članku. V primeru, da ne izberemo parametra izključnosti, dopuščamo presečnost kolokatorjev, določimo pa željeno razliko v vrednosti ključnosti kolokacije med korpusoma.

gre za pridevnik *razpadel*, ki mu je napačno dodeljena glagolska oznaka.

Zaradi filtriranja na podlagi mere *razp* smo izpustili *Praznično Družino* (posebna izdaja revije Družina), *Janševo družino* pa tudi <celje, družina>, ki izhaja iz napačne lematizacije kolokatorja *cel* v kolokaciji *cela družina*. Predvsem posamezni akterji, kot je *Janša*, se pojavljajo v različnih kolokacijah korpusa in jih je v našem primeru smiselno izpustiti.

Z ohlapnejšimi pogoji pojavitev dobimo veliko večji nabor kolokacij, npr. pri edinem pogoju logDice>3 so poleg že omenjenih kolokacij izluščene tudi:

- *raznospolna družina, partnerjeva družina*
- *črkovna družina* (grafika), *modelna družina* (avtomobilizem), *imenska družina*
- *razdirati družino*
- *časopis Družina*, *priloga Družina*, *Prehrana družine*
- *grožnja družini*

Več tovrstnih kandidatov se nahaja v enem samem viru ali pa so deli večbesednih zvez (npr. *grožnja družini* je del niza *grožnja družini Janša*). V nadaljevanju bi bilo smiselno nadgraditi pristop še z možnostjo filtriranja zadetkov, ki se pojavljajo le pri enem samem viru oz. domeni.

V primeru, da opustimo vse omejitve (logDice=0, min. frekvenca=1), je kolokacij, v katerih nastopa lema *družina* in se pojavljajo izključno v korpusu Janes, kar 180. Nekateri izmed naštetih pridevnikov, ki določajo družino so *homoseksualna*, *oligarhična*, *čefurska*, *gejevska*, zelo splošni kolokatorji so *super* in *ok,* najdemo pa tudi veliko pogostih besed, ki z lemo ne tvorijo sintaktičnih in semantičnih kolokacij (npr. *potem*, *kdo*, *pač* in *sploh*).

Podrobneje smo si ogledali primer kolokacij besede *družina.* V Tabeli 4 navajamo še kolokatorje petih drugih lem, ki so bili izluščeni z najstrožjimi pogoji (nastavitve v zadnji vrstici Tabele 3). Kot vidimo, je veliko neformalnega izrazja, vezanega na sodobne tematike (*zajebati*, *pokrasti državo*), vidimo pa tudi probleme avtomatskega luščenja kolokacij iz uporabniških spletnih besedil, saj je npr. odsotnost strešic pri zapisu vzrok za vrsto izluščenih kolokacij (npr. *smetišče zgodovine* je pogost frazem tudi v korpusu Kres, vendar ga zapis brez strešic prikaže kot specifičnega za spletno slovenščino). Med primeri vidimo tudi nestandardni zapis okrajšav brez ločil (*ang jezik*), luščenje lem na podlagi napačne lematizacije[8] (*tuja* namesto *tuj* v kolokaciji *tuj jezik*; <foto, mama> iz napačne lematizacije besede *foot,* ki ga vsebuje ime *Big foot mama*, <kurac, mama> pa iz ekspresivnega izraza *poln kurac mam* oz. *koji kurac mamo*). Korpus uporabniških spletnih vsebin je tudi bogat vir za preučevanje idiomatike (par <jezik, muca> iz Tabele 4 izvira iz izraza *muca jezik papala* (oz. *popapala, popapcala* ali celo *muca jezik papne*).

Nekatere pogoste napake predprocesiranja bi bilo mogoče odpraviti z izboljšanjem orodij za lematizacijo (učenje nad večjim označenim spletnim korpusom, boljša standardizacija), postopek rediakritizacije za pripis izpuščenih strešic.

Lema	Kolokatorji
država	vitek, zadolžiti, rušiti, pravično, ugrabiti, koruptiven, zavoziti, gnili, pokrasti, ugrabljen, zajebati
jezik	muca, eksotičen, tuja, ang, venetski, šparati
mama	foto, yo, platišče, mreža, odkar, doječ, vreme, kurac, feltna
moški	hetera, napasti, feminizacija, kastracija
zgodovina	servisen, preverljiv, smetisce, spisati, retuširanje, brisanje, odpasti

Tabela 4: Izluščeni kolokatorji za izbor petih lem.

4 Zaključek in nadaljnje delo

V prispevku smo prikazali metodo luščenja kolokacij, ki se pojavljajo le v izbranem korpusu (v našem primeru korpusu uporabniških spletnih vsebin) in ne v referenčnem korpusu. Z izdelanim orodjem za hiter dostop do kolokacij posameznega korpusa (API za dostop do korpusa na lokalni instalaciji orodja SketchEngine) lahko izvozimo kolokacije poljubnih lem. Nato seznama kolokatorjev vsake leme v različnih korpusih med seboj križamo in ohranimo le tiste kolokacijske kandidate, ki se pojavljajo izključno v specifičnem korpusu.

Z različnimi strožjimi nastavitvami parametrov frekvence besed, mere kolokabilnosti in mere specifičnosti kolokatorja na podlagi razpršenosti po seznamih (*razp*) smo iz prvotnega seznama z nad 11.000 kolokacijskimi kandidati nabor skrčili na cca. 1.600 specifičnih kolokacij. Metoda je uporabna za analizo diskurza, leksikografske naloge (sprotno dopolnjevanje drugih s specifičnimi vsebinami) ali pri poučevanju slovenščine kot tujega jezika, kjer je potrebno poznavanje in ločevanje formalnih in neformalnih načinov izražanja.

V nadaljnjem delu je potrebno predvsem izboljšati sama orodja za predprocesiranje (lematizacija), rediakritizacija pa bi pomagala luščiti bolj relevantne leksikalne specifike. S predstavljeno metodo bomo izluščili sezname kolokatorjev tudi za druge besedne vrste, zanimiva pa bi bila tudi primerjava z govornim korpusom.

5 Zahvala

Za implementacijo API vmesnika se zahvaljujem Borutu Lesjaku. Raziskava je bila opravljena v okviru projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014-2017), ki ga financira ARRS.

[8] Evalvacije v tem prispevku nismo naredili, smo pa v prispevku Pollak (2015) ocenili, da je več kot 35 % izluščenih kandidatov neprimernih zaradi predprocesiranja in sestave korpusa.

6 Literatura

Morton Benson. 1989. The structure of collocational dictionary. *The International Journal of Lexicography*, 2: 1–14.

Sabine Bartsch. 2004. *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Tübingen, Verlag Gunter Narr.

Miriam Buendía Castro in Pamela Faber. 2014. Collocation Dictionaries: A Comparative Analysis. *MonTI. Monografías de Traducción e Interpretación* 6: 203–235.

Anthony. P. Cowie. 1994. *Phraseology. Encyclopedia of Language and Linguistics* (6. zv). Oxford in New York. 3168–3171.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297–302.

John R. Firth. 1957. Modes of Meaning. Frank R. Palmer (ur.): *Papers in Linguistics 1934-51*, str. 190–215. London: Oxford University Press.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešič. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: *Zbornik 9. konference Jezikovne tehnologije*, str. 56–61.

Polona Gantar in Simon Krek. 2011. Slovene Lexical Database. V: *Zbornik 6. konference Natural language processing, multilinguality*, str. 72–80.

Laurie Gerber in Jin Yang. 1997. Systran MT dictionary development. V: *Proceedings of Past, Present, and Future: Machine Translation Summit 6,* str. 211–218.

Vojko Gorjanc in Primož Jurko. 2004. Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49(3/4): 49–62 .

Michael Alexander Kirkwood Halliday. 1966. Lexis as a Linguistic Level. *In Memory of F. R. Firth*. Longman.

Kjellmer, Göran. 1987. Aspects of English Collocations. *Corpus linguistics and Beyond*. Amsterdam, Atlanta: Rodopi.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. V: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand in Ladislav Zgusta (ur.): Wörterbücher (3 zvezki). Berlin: Walter de Gruyter.

Iztok Kosem, Polona Gantar in Simon Krek. 2013. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0* 1(2): 139–164.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: *Proceedings of EURALEX* 2004, str. 105–116.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, FDV.

Nataša Logar Berginc, Polona Gantar in Iztok Kosem. 2014. Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, str. 41–61.

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing.

Maria V. Noblia. 1998. The Computer-Mediated Communication: A New Way of Understanding The Language. V: *Proceedings of Internet Research and Information for Social Scientists Conference*, str. 10–12.

Adriane Orenha-Ottaiano. 2012. English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Language and Culture* 34 (2): 241–251.

Senja Pollak in Špela Arhar Holdt. 2015. Identifying Corpus-specific Collocations: The Case of Spoken Slovene. V: *Proceedings of 8th International Conference on Natural Language Processing, Corpus Linguistics, Lexicography* (Slovko 2015). RAM-Verlag, str. 117–125.

Senja Pollak. 2015. Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. *Zbornik 34. simpozija Obdobja. Slovnica in slovar – aktualni jezikovni opis.*

Ina Rösiger, Johannes Schäfer, Tanja George, Simon Tannert, Ulrich Heid in Michael Dorna. 2015. Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. V: *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, str. 486-503

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. V: *Proceedings of Recent Advances in Slavonic Natural Language Processing*, str. 6–9.

Violeta Seretan. 2015. Multi-Word Expressions in User-Generated Content: How Many and How Well Translated? Evidence from a Post-editing Experiment. V: *Proceedings of the Second Workshop on Multi-word Units in Machine Translation and Translation Technology (*MUMTTT 2015), Malaga, Spain.

Dirk Siepmann. 2005. Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects. *International Journal of Lexicography* 18: 409–443.

John Sinclair. 1966. Beginning the Study of Lexis. *In Memory of F. R. Firth*. London: Longman.

John Sinclair. 1991. *Corpus Concordance Collocation*. Oxford University Press.

Karen Spärck Jones. 1972 (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1): 11–22. Ponatis: *Journal of Documentation* 60 (5): 493–502.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2014. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation* 47(4): 1031-1048

Špela Vintar. 2010. Bilingual term recognition revisited. *Terminology* 16(2): 141–158.

Ning Yu. 2014. Sentiment analysis in UGC. Marie-Francine Moens, Juanzi Li, Tat-Seng Chua (ur.): *Mining User Generated Content*, str. 43–66. CRC Press. Taylor and Francis book.